

# Learning Through Reinforcement for $N$ -Person Repeated Constrained Games

Alexander S. Poznyak, *Member, IEEE*, and Kaddour Najim

**Abstract**—The design and analysis of an adaptive strategy for  $N$ -person averaged constrained stochastic repeated game are addressed. Each player is modeled by a stochastic variable-structure learning automaton. Some constraints are imposed on some functions of the probabilities governing the selection of the player's actions. After each stage, the payoff to each player as well as the constraints are random variables. No information concerning the parameters of the game is *a priori* available. The “diagonal concavity” conditions are assumed to be fulfilled to guarantee the existence and uniqueness of the Nash equilibrium. The suggested adaptive strategy which uses only the current realizations (outcomes and constraints) of the game, is based on the Bush–Mosteller reinforcement scheme in connection with a normalization procedure. The Lagrange multipliers approach with a regularization is used. The asymptotic properties of this algorithm are analyzed. Simulation results illustrate the feasibility and the performance of this adaptive strategy.

**Index Terms**—Adaptive strategies, learning automata (LA), reinforcement learning, repeated game.

## I. INTRODUCTION

GAMES THEORY operating with data of stochastic nature has seen a tremendous growth in the last decades. It began as a branch of applied mathematics, but later, when most studies were directed to games played in a stochastic environment or, to games with imperfect information, it was naturally treated as a part of Markov decision processes theory [9].

Two models of players are more frequently considered in the literature.

- 1) *Dynamic model* given by a Markov chain characterized by sets (maybe, finite, or infinite) of states and actions (several actions are allowed to be applied at each state) with the corresponding transition (in general, probabilistic) rules. This model is the main representation studied in the framework of *stochastic game theory* [6]–[8], [11], [13]. Therefore, constrained Markov chains games without learning have been studied in [1], [2], [41]. Repeated games with learning have been considered in [3], [10], [14], [22], [27], and [33]. The comprehensive review on learning in games can be found in [21].

- 2) *Static model* representing the simplest case of a finite Markov chain with a single state, that is, there are no internal dynamic effects and the only actions selected by the participants are responsible for the further behavior of each player; these models are considered within the *LA paradigm* [5], [12], [15], [16], [17], [18], [20], [23], [24], [25], [26], [28], [31], [33], [34], [40]. Recently, in [36], Bush–Mosteller reinforcement learning for zero-sum repeated game without any additional constraints has been presented. Here, we extend this approach for  $N$ -persons constraint repeated games using this paradigm.

## A. Motivation

Few papers and rigorous results, dealing with learning (for static or dynamic players model) in constrained repeated stochastic games, are actually available. The fundamental paper of Rozen [39] concerns the uniqueness problem for Nash equilibrium but with no stochastic aspect in the model. In [2], constrained stochastic games are considered but there is no learning at all. Several games with learning were considered in [19], [24], and [28], but they do not deal with situations involving constrained games. The spread of Rozen's ideas to the class of strictly diagonal concave games with reinforcement automata learning was realized in [30] (see also [31]), but the corresponding learning procedure turned out to be very complex since it included a projection operation requiring the resolution of a quadratic programming problem at every step of the game and for each player. The rigorous description of the class of the repeated games having a unique equilibrium point as well as the analysis of the convergence and convergence rate of learning for different reinforcement schemes remains an open problem that motivates partially the presented study. There are also several technical and economical situations arising in practice and generating an interest to such an investigation. One such example is given below.

*Example 1 (Pollution-Reduction Tax Game Model):* In the situation where several firms contribute to the emission of a certain pollutant, the government can detect only the combined effect. Suppose that each firm can act producing only three types of polluting emission: 1) *D*(irty); 2) *M*(iddle); and 3) *C*(lean). A *D*-production way evidently implies less production costs, a better prize product, and more profit, but the government will impose a high tax on such firm and, in addition, the public opinion may turn against it and provoke the “net effect,” e.g., through legislative measures, leading to a possible bankruptcy (all of this may occur with certain probability depending on the society's reaction). That is why the firms using a *D*-action will

Manuscript received November 15, 2001; revised March 1, 2002. This paper was recommended by Guest Editors M. S. Obaidat, G. I. Papadimitriou, and A. S. Pomportsis.

A. S. Poznyak is with the Department of Control Automatics, CINVESTAV-IPN, CP-07300 México City, México (e-mail: apoznyak@ctrl.cinvestav.mx).

K. Najim is with the Process Control Laboratory, ENSIGC., Chemin de la loge, 31078 Toulouse, France (e-mail: kaddour.najim@ensigct.fr).

Publisher Item Identifier S 1083-4419(02)06467-1.

be banned in order to keep an environment sustainable. The constraints may be related to the upper bound of an admissible financial support which a firm may realize to maintain “purifying equipments” in working conditions. Of course, each firm has a short-run initiative to make enough profit and obtain reasonable payoff, but, on the other hand, it definitely wishes to survive in the long run. That is why the average reward criterion, describing the average payoffs of each participant, is naturally applied. This is a very constrained game situation since all features characterizing the game are presented: 1) *rules* specifying what a participant can and cannot do; 2) *actions* which each player can select from a given number of alternatives; 3) *outcomes (payoffs)* corresponding to each applied action; and 4) *strategic interdependence* characterizing the fact that individual outcome depends on the actions chosen by each of the individual participants and on the probabilistic properties of environment.

### B. Discussion of Basic Assumptions and Restrictions

The most important assumptions (and, as a result, restrictions) concerning the game considered in this paper are as follows.

- 1) This is the multiperson repeated game where the behavior of each player is the static model given by a *finite state automaton* with “time-variable structure” (at each stage of the game, the distribution of the selected control actions can vary according to a reinforcement procedure).
- 2) If some controls are realized by players at a current stage, then each participant immediately gets the information on the realized values of his payoff and constraint functions which are individual for each player. The necessary information (realizations of both payoffs and constraints) is obtained during the course of the game. These realized payoffs and constraints (called *realizations*) are random variables uniformly bounded (nonobligatory binary) with probability one and having constant first (conditional) moments which are assumed to be *a priori* unknown. In other words, the game is one of incomplete information.
- 3) Each player has *his own set of constraints* (constrained functions) which may be dependent on the selected control actions of other participants (the strategic interdependence).
- 4) All participants accept some sort of agreement to use the same reinforcement learning procedure (in this case, the Bush–Mosteller scheme) *to maximize their average payoff within the given constraints* and not to change it to another one during all long-range time of the game. Any other “cooperations” during the game are prohibited. Only the parameters of this procedure can be modified by a player during the game, but not the *fixed structure* of the reinforcement procedure.

Therefore, this paper is concerned with the development of a new learning algorithm for  $N$ -person constrained repeated game with unknown expected payoff and constraints.

### C. Structure of the Paper

Some preliminary issues relevant to  $N$ -person repeated constraint game are reported in Section II. Mixed strategies and

the Nash equilibrium are discussed in Section III. The setting problem is given in Section IV. Then, the existence of the Nash equilibrium and uniqueness of optimal policies are discussed. The diagonal concavity condition and the structure of the corresponding games are studied in Section VII. Next, the uniqueness of the equilibrium point is investigated. The learning algorithm and its convergence analysis are presented in the following sections. Finally, a simulation example is presented and some conclusions end this study.

## II. $N$ -PERSON REPEATED CONSTRAINED GAME

Let  $(\Omega, \mathcal{F}, P)$  be a probability space, where  $\Omega$  is the sample space;  $\mathcal{F}$  is a minimal  $\sigma$ -algebra of subsets of  $\Omega$ ; and  $P$  is a probability measure on  $(\Omega, \mathcal{F})$ . The symbol  $\omega$  denotes the canonical point (event) in  $\Omega$ . All subsequent random variables will be defined in this space.

### A. Player Model as Learning Automaton

Each player is modeled by a *stochastic variable-structure learning automaton* which consists of a simple Markov chain containing only one state (memoryless or static system) [25], [34]. A stochastic automaton operating in a random environment (medium) is an adaptive discrete machine described by the tuple  $\{\Xi, U^k, \{\xi_n^k\}, \{u_n^k\}, \{p_n^k\}, T^k\}$  where  $\Xi$  is the automaton input bounded set;  $U^k$  denotes the set  $\{u^k(1), u^k(2), \dots, u^k(N_k)\}$  ( $k = \overline{1, N}$ ) of actions of the automaton;  $\{\xi_n^k\}, \{\eta_n^{k,l}\}$  is a sequence of automaton inputs [payoffs and constraints  $(\xi_n^k, \eta_n^{k,l}) \in \Xi, l = \overline{1, M_k}$ ] provided by the game mechanism in a binary ( $P$ -model environment) or continuous ( $S$ -model environment) form;  $\{u_n^k\}$  is a sequence of automaton outputs (actions); and  $p_n^k = [p_n^k(1), p_n^k(2), \dots, p_n^k(N_k)]^T$  is the conditional probability distribution at time  $n$

$$p_n^k(i) = P\{\omega \in \Omega: u_n^k = u^k(i) / \mathcal{F}_{n-1}\}, \quad \sum_{i=1}^{N_k} p_n^k(i) = 1$$

where  $\mathcal{F}_n = \sigma(u_1^k, p_1^k, \xi_1^k, \eta_1^{k,l}, \dots; u_n^k, p_n^k, \xi_n^k, \eta_n^{k,l}; l = \overline{1, M_k}, k = \overline{1, N})$  is the minimal  $\sigma$ -algebra generated by the corresponding events ( $\mathcal{F}_n \subseteq \mathcal{F}$ );  $T^k = T_n^k$  represents the *reinforcement scheme* (updating scheme) which changes the probability vector  $p_n^k$  to  $p_{n+1}^k$ , that is

$$p_{n+1}^k = p_n^k + \gamma_n^k T_n^k$$

$$T_n^k = T_n^k(p_n^k, \{\xi_t^k\}_{t=1, \dots, n}, \{\eta_t^{k,l}\}_{t=1, \dots, n}, \{u_t^k\}_{t=1, \dots, n}) \quad (1)$$

$p_1^k(i) > 0 \forall i = 1, \dots, N_k$  where  $\gamma_n^k$  is a scalar correction factor and the vector  $T_n^k(\cdot) = [T_n^{k,1}(\cdot), \dots, T_n^{k,N_k}(\cdot)]^T$  satisfies the following conditions (for preserving probability measure):  $\forall n, \forall i = 1, \dots, N$

$$\sum_{i=1}^N T_n^{k,i}(\cdot) = 0, \quad p_n^k(i) + \gamma_n^k T_n^{k,i}(\cdot) \in [0, 1].$$

This is the heart of the learning automaton (LA).

### B. How the Game Is Played

This game is played in the following way. According to the probability distributions (mixed strategy)  $p_n^k$ , at each stage  $n$ , simultaneously and independently (we consider a noncooperative game in which each player has only its own payoff realization  $\xi_n^k = \xi_n^k(\omega)$  and the realizations  $\eta_n^{k,l} = \eta_n^{k,l}(\omega)$  of the constraints), each player chooses randomly an action, for instance  $u^k(i_k)$  ( $i_k = \overline{1, N_k}$ ,  $k = \overline{1, N}$ ).

*Remark 2:* The technique used to select an action  $u^k(i_k)$  among  $N_k$  actions is based on the generation of uniformly distributed random variable  $z$  (any specific machine routine: RAND, etc., can be used to carry out the random variable  $z$ ). The  $k$ th automaton (player) chooses the action  $u^k(i_k)$  such that  $i_k$  is equal the least value of  $i$ , verifying the following constraint:  $\sum_{j=1}^i p_n^k(j) \geq z \in [0, 1]$ .

These actions produce the random reactions

$$\begin{cases} \xi_n^k := (\xi_n^{k,1}, \dots, \xi_n^{k,N_k})^T \in R^{N_k} \\ \xi_n^k = \xi^k(\omega; u_n^k = u^k(i_k), k = \overline{1, N}) \\ \eta_n^{k,l} = (\eta_n^{k,l,1}, \dots, \eta_n^{k,l,M_k})^T \in R^{M_k} \\ \eta_n^{k,l} = \eta^{k,l}(\omega; u_n^k = u^k(i_k)). \end{cases}$$

The  $k$ th random component  $\xi_n^k$  represents the payoff (Borel function) earned by the  $k$ th player, and  $\eta_n^{k,l}$  is the random realization of the  $l$ th constraint for this player. To obtain reasonable generality without excessive complexity, the following assumptions will be in force throughout this paper:

*Assumption A1:* The conditional expectations of  $\xi_n^k$  and  $\eta_n^{k,l}$  are independent of the history of the game (the game rules are assumed to be unchangeable), that is

$$\begin{aligned} E\{\xi_n^k | u_n^k = u^k(i_k) \wedge F_{n-1}, k = \overline{1, N}\} \\ := \int_{\omega \in \Omega} \xi^k(\omega; u_n^k = u^k(i_k)) dP(\omega) \stackrel{a.s.}{=} v_{i_1, \dots, i_N}^k \\ E\{\eta_n^{k,l} | u_n^k = u^k(i_k) \wedge F_{n-1}, k = \overline{1, N}\} \\ := \int_{\omega \in \Omega} \eta^{k,l}(\omega; u_n^k = u^k(i_k), k = \overline{1, N}) dP(\omega) \stackrel{a.s.}{=} q_{i_1, \dots, i_N}^{k,l}. \end{aligned} \quad (2)$$

*Assumption A2:* For any realized actions  $u_n^k$  at current stage  $n$  of the play, the absolute values of  $\xi_n^k$  and  $\eta_n^{k,l}$  are uniformly (on  $\omega$  and  $n$ ) bounded with probability one, that is

$$\begin{aligned} |\xi^k(\omega; u_n^k = u^k(i_k), k = \overline{1, N})| \stackrel{a.s.}{\leq} \sigma_{\xi, k}^+ < \infty \\ |\eta^l(\omega; u_n^k = u^k(i_k), k = \overline{1, N})| \stackrel{a.s.}{\leq} \sigma_{\eta, k, l}^+ < \infty \end{aligned} \quad (3)$$

where  $|\cdot|$  stands for the absolute value. For each player, the average payoff and constraints form the collection of  $M_k + 1$  tensors  $V^k = [v_{i_1, \dots, i_N}^k]$  ( $i_k = \overline{1, N_k}$ ,  $k = \overline{1, N}$ ) and  $Q^{k,l} = [q_{i_1, \dots, i_N}^{k,l}]$  ( $l = \overline{1, M_k}$ ) which are assumed to be *a priori* unknown. Then, each player changes his mixed strategy according to the accepted reinforcement ( $p_n^k \xrightarrow{T_n^k} p_{n+1}^k$ ) trying for large  $n$  to maximize his individual averaged payoff  $\liminf_{n \rightarrow \infty} \Phi_n^k$  where

$$\Phi_n^k := n^{-1} \sum_{t=1}^n \xi_t^k \quad (4)$$

maintaining (with probability one) the constraints

$$\liminf_{n \rightarrow \infty} \Psi_n^{k,l} \leq b^{k,l}, \quad \Psi_n^{k,l} := n^{-1} \sum_{t=1}^n \eta_t^{k,l}. \quad (5)$$

The next definition concerns the *expected payoff*.

*Definition 3:* At stage  $n$ , for the  $k$ th player of the considered game, and for any  $F_{n-1}$ -measurable conditional probability distribution  $(p_n^1, \dots, p_n^N)$ , the Borel function

$$V^k(p_n^1, \dots, p_n^N) := \sum_{i_1=1}^{N_1} \dots \sum_{i_N=1}^{N_N} v_{i_1, \dots, i_N}^k \prod_{s=1}^N p_n^s(i_s) \quad (6)$$

is said to be the *expected payoff* and

$$Q^{k,l}(p_n^1, \dots, p_n^N) := \sum_{i_1=1}^{N_1} \dots \sum_{i_N=1}^{N_N} q_{i_1, \dots, i_N}^{k,l} \prod_{s=1}^N p_n^s(i_s) \quad (7)$$

corresponds to the *expected constraints*.

We also define the *averaged expected payoff* and *constraints* as follows:

$$\begin{aligned} V_n^k &:= n^{-1} \sum_{t=1}^n V^k(p_t^1, \dots, p_t^N) \\ Q_n^{k,l} &:= n^{-1} \sum_{t=1}^n Q^{k,l}(p_t^1, \dots, p_t^N). \end{aligned} \quad (8)$$

The following lemma states the asymptotic equivalence between  $(\Phi_n^k, \Psi_n^{k,l})$  and  $(V_n^k, Q_n^{k,l})$ .

*Lemma 4:* Under *Assumption A1*, for any conditional distributions  $\{p_n^1, \dots, p_n^N\}$ , it follows:

$$\Phi_n^k \stackrel{a.s.}{=} V_n^k + o_\omega(n^{-(1/2)}), \quad \Psi_n^{k,l} \stackrel{a.s.}{=} Q_n^{k,l} + o_\omega(n^{-(1/2)}).$$

*Proof:* The proof follows immediately from [35, Lemma 5, App. A] for  $v_t := \eta_t = n^{1-\varepsilon}$ ,  $\alpha_n := n^{-1}$ , and  $\beta_n := n^{-2}$ . ■

### III. MIXED STRATEGIES AND NASH EQUILIBRIUM

#### A. Randomized Strategies

The *randomized (or mixed) strategy* of the  $k$ th player is any sequence of vectors  $D^k = \{p_n^k\}$  ( $k = \overline{1, N}$ ) with  $F_{n-1}$ -measurable components belonging to the simplex  $S_{\varepsilon=0}^{N_k}$  (for the regularization purpose we will use  $\varepsilon = \varepsilon_n > 0$ ), that is

$$\begin{aligned} p_n^k &\in S_{\varepsilon=0}^{N_k} \\ &:= \left\{ p_n^k \in R^{N_k} : p_n^k(i) \geq \varepsilon \geq 0, \sum_{i=1}^{N_k} p_n^k(i) = 1 \right\}. \end{aligned} \quad (9)$$

#### B. Noncooperative Equilibrium

*Definition 5:* The strategies  $\overline{D}^1, \dots, \overline{D}^N$  are said to be the *noncooperative equilibrium strategies* (in the Nash sense) if:

1) they are admissible, i.e., for any  $k = \overline{1, N}$ ,  $l = \overline{1, M_k}$

$$\max_{p_i^1, s=1, N} \limsup_{n \rightarrow \infty} Q_n^{k,l} := \tilde{Q}(\overline{D}^1, \dots, \overline{D}^N) \leq b^{k,l}$$

where  $b^{k,l} \in R$  are *a priori* given;

2) for any integer  $k$  and any admissible strategy  $D^k$

$$\begin{aligned} \tilde{V}^k(\bar{D}^1, \dots, \bar{D}^N) \\ &:= \min_{p_1^s, s=\overline{1, N}} \liminf_{n \rightarrow \infty} V_n^k \\ &\stackrel{a.s.}{\geq} \tilde{V}(\bar{D}^1, \dots, \bar{D}^{k-1}, D^k, \bar{D}^{k+1}, \dots, \bar{D}^N) \end{aligned} \quad (10)$$

where the minimization is done over all the initial probability distributions  $p_1^s, s = \overline{1, N}$ .

### C. Equilibrium Within the Subclass of Stationary Mixed Strategies

Consider now the subclass of stationary mixed strategies  $D^k = \{p^k\}$  ( $k = \overline{1, N}$ ).

**Definition 6:** The point  $(\bar{p}^1, \dots, \bar{p}^N)$  is said to be an *equilibrium point* of the given  $N$ -person game within the class of stationary strategies if, for each  $k = \overline{1, N}$

$$\begin{aligned} V^k(\bar{p}^1, \dots, \bar{p}^N) &= \max_{p^k \in R^k} V^k(\bar{p}^1, \dots, \bar{p}^{k-1}, p^k, \bar{p}^{k+1}, \dots, \bar{p}^N) \\ R^k &:= S_0^{N_k} \bigcap_{l=1}^{M_k} \{p^k : Q^{k,l}(\bar{p}^1, \dots, \bar{p}^{k-1}, \\ &\quad p^k, \bar{p}^{k+1}, \dots, \bar{p}^N) \leq b^{k,l}\}. \end{aligned} \quad (11)$$

At this point, no player can increase his payoff by a unilateral change in his strategy.

**Remark 7:** Evidently, the set of all equilibrium strategies  $\bar{D}^k$  ( $k = \overline{1, N}$ ) contains the subset of *admissible stationary strategies*  $\{\bar{p}^k\}$  for any  $k = \overline{1, N}, l = \overline{1, M_k}$  and for any  $p^k \in S_{\varepsilon=0}^{N_k}$  realizing the inequalities

$$\begin{aligned} V^k(\bar{p}^1, \dots, \bar{p}^N) &\geq V^k(\bar{p}^1, \dots, \bar{p}^{k-1}, p^k, \bar{p}^{k+1}, \dots, \bar{p}^N) \\ Q^{k,l}(\bar{p}^1, \dots, \bar{p}^{k-1}, p^k, \bar{p}^{k+1}, \dots, \bar{p}^N) &\geq b^{k,l}. \end{aligned} \quad (12)$$

This fact follows from: 1) the continuity property of the function  $V^k$  (8) and  $Q^{k,l}$  (8); 2) the compactness of the simplexes  $S_{\varepsilon=0}^{N_k}$ ; and 3) the Nash theorem [29, Th. 1, p. 268].

**Remark 8:** Notice that Pareto payoffs can be achieved as Nash equilibria (see, for example, [9] and [11, Pt. II]).

## IV. GAME AS REINFORCEMENT: PROBLEM SETTING

Now we are ready to formulate the  $N$ -person repeated constrained game problem with *a priori* unknown average payoffs and constraints. Based on current information, generate randomized (mixed) admissible strategies  $\{p_n^k\}$  ( $k = \overline{1, N}$ ) in order to achieve a Nash equilibrium realizable within the subclass of stationary strategies.

To achieve this objective, let us first emphasize the following fact. According to Nash theorem [29], the set of stationary distributions  $(\bar{p}^1, \dots, \bar{p}^N)$ , satisfying (12) may contain more than one element. Consider one of them and show that if an admissible mixed strategy  $D^k$  ( $k = \overline{1, N}$ ) converges to such probability distribution  $(\bar{p}^1, \dots, \bar{p}^N)$ , then the associated random functions  $\Phi_n^k$  and  $\Psi_n^{k,l}$  (4) also converge (with the same rate) to the corresponding average values  $V^k(\bar{p}^1, \dots, \bar{p}^N)$  and  $Q^{k,l}(\bar{p}^1, \dots, \bar{p}^N)$ , respectively.

**Lemma 9:** If, under *Assumptions A1* and *A2*, there exists a positive parameter  $\tau$  such that

$$\limsup_{n \rightarrow \infty} n^\tau E \left\{ \sum_{k=1}^N \|p_n^k - \bar{p}^k\|^2 \right\} < \infty$$

that is, the strategy  $D^k$  ( $k = \overline{1, N}$ ) is asymptotically stationary realizing a Nash equilibrium, then

$$\begin{aligned} \limsup_{n \rightarrow \infty} n^\tau E \left\{ \sum_{k=1}^N |\Phi_n^k - V^k(\bar{p}^1, \dots, \bar{p}^N)|^2 \right\} &< \infty \\ \limsup_{n \rightarrow \infty} n^\tau E \left\{ \sum_{l=1}^{M_k} |\Psi_n^{k,l} - Q^{k,l}(\bar{p}^1, \dots, \bar{p}^N)|^2 \right\} &< \infty. \end{aligned} \quad (13)$$

**Proof:** Consider the functions

$$\begin{aligned} d_n^{1,k}(\{u_n^k\}, \{p_n^k\}) &:= \sum_{k=1}^N [\Phi_n^k - V^k(\bar{p}^1, \dots, \bar{p}^N)]^2 \\ d_n^{2,k}(\{u_n^k\}, \{p_n^k\}) &:= \sum_{l=1}^{M_k} [\Psi_n^{k,l} - Q^{k,l}(\bar{p}^1, \dots, \bar{p}^N)]^2. \end{aligned}$$

Taking into account that the functions  $V^k(\bar{p}^1, \dots, \bar{p}^N)$  and  $Q^{k,l}(\bar{p}^1, \dots, \bar{p}^N)$  are smooth enough and satisfy the Lipschitz condition, one has (for  $r = 1, 2$ )

$$\begin{aligned} E \{ d_n^{r,k}(\{u_n^k\}, \{p_n^k\}) | \hat{F}_{n-1} \} &\leq C \sum_{k=1}^N \|p_n^k - \bar{p}^k\| / n \\ &\quad + \left(1 - \frac{1}{n}\right) d_{n-1}^{r,k}(\{u_{n-1}^k\}, \{p_{n-1}^k\}) + \frac{C}{n^2}. \end{aligned}$$

In view of [34, Lemma A.3-1, App. A], it follows that  $d_n^{r,k}(\{u_n^k\}, \{p_n^k\}) \xrightarrow{a.s.} 0$ ,  $r = 1, 2$ . Using [34, Lemma A.3-22, App. A] and considering the mathematical expectation of both sides of the previous inequality, we obtain the desired result (13). ■

Therefore, if we construct an asymptotically stationary mixed strategy  $\{p_n^k\}$  converging to a stationary distribution  $(\bar{p}^1, \dots, \bar{p}^N)$  realizing a Nash equilibrium, we will be able to attain the main aim of the game. However, to do this in a rigorous manner, first, the problems related to the *existence* and the *uniqueness* of the Nash equilibrium (within stationary strategies) should be solved. Regarding the uniqueness of the finite matrix games, one can consider the uniqueness of the optimal policy for each player as well as the Nash equilibrium point. The next sections deal with these problems.

## V. EXISTENCE OF NASH EQUILIBRIUM

### A. Slater's Condition

Assume that the given constraints satisfy the Slater's condition providing the Kuhn–Tucker constraint qualification [4] (to justify the correctness of Lagrange multipliers implementation and the existence of nonempty set of admissible strategies):

*Assumption A3:* There exist vectors  $(\hat{p}_n^1, \dots, \hat{p}_n^N)$  such that for all  $k = \overline{1, N}$  and  $l = \overline{1, M_k}$  the following inequalities strictly hold:

$$Q^{k,l}(\hat{p}_n^1, \dots, \hat{p}_n^N) = \sum_{i_1=1}^{N_1} \dots \sum_{i_N=1}^{N_N} q_{i_1, \dots, i_N}^{l,k} \prod_{s=1}^N \hat{p}_n^s(i_s) < b^{k,l}. \quad (14)$$

### B. Main Result on Existence

*Lemma 10:* An equilibrium point within the class of stationary strategies exists for any  $N$ -person game satisfying **(A1)** and **(A3)**.

*Proof:* It follows straight from [39, Th. 1]. Indeed, by **(A1)** the function

$$\rho(p, q) := \sum_{k=1}^N V^k(p^1, \dots, q^k, \dots, p^N) \quad (15)$$

is well defined for any  $p := (p^1, \dots, p^N)$  and  $q := (q^1, \dots, q^N)$  from  $R = R^1 \times \dots \times R^N$  and, by (8), is continuous in  $p$  and  $q$  and linear (and, hence, concave) in  $q$  for any fixed  $p$  within  $R$ . Notice that  $R$  is a convex [nonempty by **(A2)** (14)] compact. Then, the point-to-set mapping  $p \in R \rightarrow \Gamma p \subset R$  given by

$$\Gamma p := \left\{ q \mid \rho(p, q) = \max_{z \in R} \rho(p, z) \right\} \quad (16)$$

is upper semicontinuous in  $R$  and, hence, by the Kakutani fixed point theorem (see, for example, [11]), there exist a point  $\bar{p} \in R$  such that  $\bar{p} \in \Gamma \bar{p}$ , that is

$$\rho(\bar{p}, \bar{p}) = \max_{z \in R} \rho(\bar{p}, z). \quad (17)$$

This fixed point  $\bar{p} \in R$  satisfies (11). Indeed, suppose that (11) is not verified, e.g., for  $k = k_0$ , there would be a point  $p^{k_0}$  such that for  $\bar{p}(k_0) := (\bar{p}^1, \dots, p^{k_0}, \dots, \bar{p}^N) \in R$  we have  $V^k(\bar{p}^1, \dots, p^{k_0}, \dots, \bar{p}^N) > V^k(\bar{p}^1, \dots, \bar{p}^k, \dots, \bar{p}^N)$  that implies  $\rho(\bar{p}, \bar{p}(k_0)) > \rho(\bar{p}, \bar{p})$  which contradicts (17). Lemma is proven. ■

## VI. PAYOFF REGULARIZATION AND UNIQUENESS OF OPTIMAL POLICIES

By the definition of (11) and multilinearity of the payoff functions  $V^k(p^1, \dots, p^N)$  given by (8), it follows that even the optimal policy for each player may be not unique. To provide the uniqueness of the optimal policies, following [31], let us introduce the so-called *regularized payoff function* as follows:

$$V_\delta^k(p^1, \dots, p^N) := V^k(p^1, \dots, p^N) - \frac{\delta}{2} \|p^k\|^2$$

$$p^k \in S_\varepsilon^{N_k}, \varepsilon \in (0, N_k^{-1}), \delta > 0. \quad (18)$$

The initially given payoffs evidently correspond to the limit case  $\varepsilon = \delta = 0$ . The function  $V_\delta^k(\bar{p}^1, \dots, p^k, \dots, \bar{p}^N)$  is strictly concave in  $p^k$  if  $\delta > 0$ , and hence, the optimal policy

$$\bar{p}^k = \arg \max_{p^k \in R_\varepsilon^k} V_\delta^k(\bar{p}^1, \dots, p^k, \dots, \bar{p}^N)$$

$$R_\varepsilon^k := S_\varepsilon^{N_k} \bigcap \{p^k: Q^{k,l}(\bar{p}^1, \dots, p^k, \dots, \bar{p}^N) \leq b^{k,l}\} \quad (19)$$

is unique in  $p^k \in R_\varepsilon^k$  for any given  $\bar{p}^{s \neq k}$ . However, this does not imply a unique equilibrium.

## VII. DIAGONAL CONCAVITY

As it is shown by Rosen [39], the nonuniqueness of the Nash equilibrium points cannot be offset by a small regularization (perturbation) term. The condition for the uniqueness of equilibria are known as *strict diagonal concavity*. Next, following [31], we show that if the given (nonregularized) matrix game is “*diagonal concave*,” then the corresponding regularized game turns out to be strictly diagonal concave that, by Rosen’s theorem [39, Th. 2], the uniqueness of the equilibria policy follows.

### A. Diagonal Concavity Property

Let us define the function

$$W_r(p, q) := (q - p) \frac{\partial}{\partial p} \rho_r(p, q) + (p - q) \frac{\partial}{\partial q} \rho_r(p, q)$$

$$\rho_r(p, q) := \sum_{k=1}^N r_k V^k(p^1, \dots, q^k, \dots, p^N). \quad (20)$$

It is easy to check that

$$W_r(p, q) = \sum_{k=1}^N r_k [V^k(p^1, \dots, q^k, \dots, p^N) - V^k(p^1, \dots, p^N) + V^k(q^1, \dots, p^k, \dots, q^N) - V^k(q^1, \dots, q^N)].$$

*Definition 11:* A matrix game is said to be *diagonal concave* if there exist positive numbers  $r_k$  such that for any  $p, q \in R$

$$W_r(p, q) \leq 0. \quad (21)$$

Several useful statements concerning the basic properties of the diagonal concave matrix games are given [31].

### B. The Set of Nash Equilibrium Points

*Lemma 12:* The set  $\mathbf{P}^*$  of all Nash equilibria points for any diagonal concave game is closed and convex.

*Proof:* The closeness follows from the continuity property of  $V^k(p^1, \dots, p^N)$ ,  $p \in R$ . To prove the convexity let us consider  $\bar{p}, \tilde{p} \in \mathbf{P}^*$  and  $\lambda \in [0, 1]$ . Define  $z := \lambda \bar{p} + (1 - \lambda) \tilde{p} \in R$ . Then, by (21) and by the optimality property (11) for any  $q \in R$ , it follows:

$$\begin{aligned} & \sum_{k=1}^N r_k [V^k(z^1, \dots, q^k, \dots, z^N) - V^k(z^1, \dots, z^N)] \\ & \leq \sum_{k=1}^N r_k [V^k(q^1, \dots, z^k, \dots, q^N) - V^k(q^1, \dots, q^N)] \\ & = \lambda \sum_{k=1}^N r_k [V^k(q^1, \dots, \bar{p}^k, \dots, q^N) - V^k(q^1, \dots, q^N)] \\ & \quad + (1 - \lambda) \sum_{k=1}^N r_k [V^k(q^1, \dots, \tilde{p}^k, \dots, q^N) - V^k(q^1, \dots, q^N)] \\ & \leq \lambda \sum_{k=1}^N r_k [V^k(\bar{p}^1, \dots, q^k, \dots, \bar{p}^N) - V^k(\bar{p}^1, \dots, \bar{p}^N)] \\ & \quad + (1 - \lambda) \sum_{k=1}^N r_k [V^k(\tilde{p}^1, \dots, q^k, \dots, \tilde{p}^N) - V^k(\tilde{p}^1, \dots, \tilde{p}^N)] \leq 0 \end{aligned}$$

that for  $q^{s \neq k} = z^s$  implies  $V^k(z^1, \dots, z^N) \geq V^k(z^1, \dots, q^k, \dots, z^N)$  and hence,  $z \in \mathbf{P}^*$ . ■

### C. Structure of Diagonal Concave Matrix Games

**Lemma 13 (First Criterion of Diagonal Concavity):** A given game is diagonal concave, if and only if there exist positive numbers  $r_k$  such that for any  $p, q \in R$

$$W_r(p, q) = 0.$$

*Proof:* This proof follows from [31, Lemma 4.4]. Sufficiency is evident. Let us prove the necessity. For any two fixed combination  $I := (i_1, \dots, i_N)$  and  $J := (j_1, \dots, j_N)$  of indexes consider the set  $T(I, J)$  of all possible pairs  $A = (a_1, \dots, a_N)$  and  $B = (b_1, \dots, b_N)$  of indexes having whether  $(a_k = i_k; b_k = j_k)$  or  $(a_k = j_k; b_k = i_k)$ . Since in view of (21) for

$$p^k = e_{i_k}^{N_k} := \left( \underbrace{0, 0, \dots, 0}_{i_k}, 1, 0, \dots, 0 \right)^T \in S_0^{N_k}$$

and

$$q^k = e_{j_k}^{N_k} := \left( \underbrace{0, 0, \dots, 0}_{j_k}, 1, 0, \dots, 0 \right)^T \in S_0^{N_k}$$

it follows that

$$\begin{aligned} W_{I, J} &:= W_r(p, q) \\ &= \sum_{k=1}^N r_k \left[ v_{i_1, \dots, i_{k-1}, j_k, i_{k+1}, \dots, i_N}^k - v_{i_1, \dots, i_N}^k \right. \\ &\quad \left. + v_{j_1, \dots, j_{k-1}, i_k, j_{k+1}, \dots, j_N}^k - v_{j_1, \dots, j_N}^k \right] \leq 0 \end{aligned}$$

and  $\sum_{(A, B) \in T(I, J)} W_{A, B} = 0$ . Since  $W_{A, B} \leq 0$  for any  $(A, B) \in T(I, J)$ , that implies that  $W_{I, J} = 0$ . In view of this and taking into account the representation  $W_r(p, q) = \sum_{I, J} W_{I, J} \prod_{k=1}^N p^k(i_k) q^k(j_k)$  we obtain  $W_r(p, q) = 0$ . Lemma is proven. ■

**Lemma 14 (Second Criterion of Diagonal Concavity):** A matrix game with  $N$  participants is diagonal concave if and only if there exist numbers  $r_k > 0$ ,  $a_{i_k}^k$ ,  $b_{i_k, j_s}^{k, s}$  and  $c_{i_1, \dots, i_N}^k$  such that an expected payoffs  $V^k$  for each player can be represented in the form

$$\begin{aligned} V^k(p^1, \dots, p^N) \\ &:= \sum_{i_k=1}^{N_k} a_{i_k}^k \prod_{k=1}^N p^k(i_k) + r_k^{-1} \sum_{s \neq k}^N \sum_{i_k=1}^{N_k} \sum_{i_s=1}^{N_s} \\ &\quad \cdot b_{i_k, j_s}^{k, s} p^k(i_k) p^s(i_s) + \sum_{i_1=1}^{N_1} \dots \sum_{i_{k-1}=1}^{N_{k-1}} \sum_{i_{k+1}=1}^{N_{k+1}} \dots \sum_{i_N=1}^{N_N} \\ &\quad \cdot c_{i_1, \dots, i_{k-1}, i_{k+1}, i_N}^k \prod_{s \neq k}^N p^s(i_s). \end{aligned} \quad (22)$$

*Proof:* Defining  $U(p, q) := \sum_{k=1}^N r_k [V^k(p^1, \dots, q^k, \dots, p^N) - V^k(p^1, \dots, p^N)]$ , from the previous lemma it follows that  $W_r(p, q) = U(p, q) + U(q, p) = 0$ . Excluding the variables  $p^k(i_k) = 1 - \sum_{s \neq k}^N p^k(i_s)$  and  $q^k(i_k) = 1 - \sum_{s \neq k}^N q^k(i_s)$  from  $U(p, q)$  we can define the function

$\tilde{U}(\tilde{p}, \tilde{q})$  of variables  $\tilde{p} = \tilde{p}^1, \dots, \tilde{p}^N$ ,  $\tilde{q} = \tilde{q}^1, \dots, \tilde{q}^N$  ( $\tilde{p}^1 := (p^k(1), \dots, p^k(N_k - 1))^T$ ) which satisfies

$$\tilde{U}(\tilde{p}, \tilde{q}) + \tilde{U}(\tilde{q}, \tilde{p}) = 0, \quad \tilde{U}(\tilde{p}, \tilde{p}) = 0. \quad (23)$$

Since the function  $\tilde{U}(\tilde{p}, \tilde{q})$  is linear on  $\tilde{p}$  (under the fixed  $\tilde{q}$ ) and vise versa, and it also does not contain the products  $p^s(i_s) q^s(i_s)$ , it follows that

$$\begin{aligned} \tilde{U}(\tilde{p}, \tilde{q}) &= u_0 + \sum_{k=1}^N \sum_{i_k=1}^{N_k-1} (\alpha_{i_k}^k p^k(i_k) - \bar{\alpha}_{i_k}^k q^k(i_k)) \\ &\quad + \sum_{k=1}^N \sum_{s \neq k}^N \sum_{i_k=1}^{N_k-1} \sum_{i_s=1}^{N_s-1} \beta_{i_k, i_s}^{k, s} p^k(i_k) q^s(i_s) \end{aligned}$$

which, together with (23), leads to following the relations:

- 1) for  $\tilde{p} = 0$ :  $u_0 = 0$ ;
- 2) for  $p^s(i_s) = \delta_{i_s, j_s}$ ,  $\tilde{p}^k = 0$  ( $k \neq s$ ):  $\alpha_{i_k}^k = \bar{\alpha}_{i_k}^k$ ;
- 3) for  $p^s(i_s) = \delta_{i_s, j_s}$ ,  $\tilde{p}^k = 0$  ( $k \neq s$ ,  $k \neq \tau$ ),  $p^\tau(i_\tau) = \delta_{i_\tau, j_\tau}$ :  $\beta_{j_k, j_\tau}^{k, \tau} + \beta_{j_\tau, j_k}^{\tau, k} = 0$ ;

which imply (22). Lemma is proven. ■

It is easy to check that the payoffs of any diagonal concave matrix game are linear combinations of the only three simple diagonal concave games: two-person zero-sum matrix games, matrix games where the payoff of each player does not depend on the strategies of other participants ("games with nature") and matrix games where the payoff of each player does not depend on his own strategies and is determined by the strategies of other participants.

## VIII. STRICTLY DIAGONAL CONCAVE GAMES AND UNIQUENESS OF EQUILIBRIUM POINT

As already mentioned above, the equilibrium probability distribution may be not unique. To avoid this indeterminacy, we shall use the regularization approach presented in [34].

### A. Strict Diagonal Concavity

The consideration of diagonal concave games, where the payoff  $V^k$  (8) is changed to  $V_\delta^k$  (18), directly leads to the following result:

$$\begin{aligned} W_r^\delta(p, q) &:= (q - p) \frac{\partial}{\partial p} \rho_r^\delta(p, q) + (p - q) \frac{\partial}{\partial q} \rho_r^\delta(p, q) \\ &= \sum_{k=1}^N r_k [V_\delta^k(p^1, \dots, q^k, \dots, p^N) \\ &\quad - V_\delta^k(p^1, \dots, p^N) + V_\delta^k(q^1, \dots, p^k, \dots, q^N) \\ &\quad - V_\delta^k(q^1, \dots, q^N)] \\ &\leq -\delta \sum_{k=1}^N r_k \|p^k - q^k\|^2 \end{aligned} \quad (24)$$

where

$$\rho_r^\delta(p, q) := \sum_{k=1}^N r_k V_\delta^k(p^1, \dots, q^k, \dots, p^N). \quad (25)$$

Here, the equality is possible iff  $p^k = q^k$  for all  $k = \overline{1, N}$ .

*Definition 15:* The games satisfying (24) are named *strictly diagonal concave*.

According to Rosen's theorem [39, Th. 2], such games have the unique equilibria point (policy). Thereafter, we will deal with the regularized payoffs where the regularized parameter tends to zero during the learning process.

### B. Regularized Lagrange Function

Consider the following regularized (augmented) Lagrange functions:

$$\begin{aligned} L_\delta^k(p^1, \dots, p^N; \lambda) &:= V_\delta^k(p^1, \dots, p^N) + \frac{\delta}{2} \|\lambda\|^2 \\ &\quad - \sum_{l=1}^{M_k} \lambda^{k,l} [Q^{k,l}(\bar{p}^1, \dots, \bar{p}^N) - b^l] \\ \lambda^k &:= (\lambda^{k,1}, \dots, \lambda^{k,M_k})^T; \quad \lambda^{k,l}, \delta > 0 \end{aligned} \quad (26)$$

where the arguments  $p^k$  belong to the  $\varepsilon$ -simplexes (9). These regularized functions are strictly concave with respect to  $p^k$ , and strictly convex with respect to  $\lambda^k$ . Hence, these regularized Lagrange functions satisfy (see [34, Lemma 1, p. 116])

$$\begin{aligned} &\sum_{k=1}^N r_k \left[ (p_n^k - \bar{p}^k(\varepsilon_n^k, \delta_n))^T \frac{\partial}{\partial p_n^k} L_\delta^k(p_n^1, \dots, p_n^N; \lambda_n) \right. \\ &\quad \left. - (\lambda_n^k - \lambda^k(\varepsilon_n, \delta_n))^T \frac{\partial}{\partial \lambda_n} L_\delta^k(p_n^1, \dots, p_n^N; \lambda_n) \right] \\ &\leq -\frac{\delta}{2} \sum_{k=1}^N r_k \left( \|p_n^k - \bar{p}^k(\varepsilon_n^k, \delta_n)\|^2 + \|\lambda_n^k - \lambda^k(\varepsilon_n, \delta_n)\|^2 \right) \end{aligned} \quad (27)$$

where the corresponding equilibrium point is denoted by  $(\bar{p}^1(\varepsilon, \delta), \dots, \bar{p}^N(\varepsilon, \delta))$  and the associated Lagrange multipliers are denoted by  $\lambda^k(\varepsilon, \delta)$ .

### C. Parametrization of the Set of Equilibrium Points

*Lemma 16:* If the sequences  $\varepsilon_n^k$  and  $\{\delta_n\}$  are such that

$$\begin{aligned} \varepsilon_n^k &\in (0, N_k^{-1}), \quad \delta_n > 0, \quad n = 1, 2, \dots \\ \lim_{n \rightarrow \infty} \delta_n &= 0, \quad \lim_{n \rightarrow \infty} \varepsilon_n \delta_n^{-1} = \vartheta := (\vartheta^1, \dots, \vartheta^N)^T \\ \varepsilon_n &:= (\varepsilon_n^1, \dots, \varepsilon_n^N)^T, \quad \vartheta^s \in [0, \infty) \end{aligned}$$

then

- 1) all the possible equilibrium mixed policies can be parametrized by the nonnegative vector parameter  $\vartheta \in R^N$  as  $(\bar{p}^1(\vartheta), \dots, \bar{p}^N(\vartheta))$  and the sequences  $(\bar{p}^1(\varepsilon_n^1, \delta_n), \dots, \bar{p}^N(\varepsilon_n^N, \delta_n))$  and  $\lambda^k(\varepsilon_n, \delta_n)$  converge, respectively, to the unique points  $(\bar{p}^1(\vartheta), \dots, \bar{p}^N(\vartheta))$  and  $\lambda^k(\vartheta)$ , i.e.

$$\sum_{k=1}^N r_k (\| \bar{p}^k(\varepsilon_n^k, \delta_n) - \bar{p}^k(\vartheta) \| + \| \lambda^k(\varepsilon_n, \delta_n) - \lambda^k(\vartheta) \|) \xrightarrow{n \rightarrow \infty} 0.$$

- 2) There exist constants  $C_i$  ( $i = 1, 2, 3$ ) such that

$$\begin{aligned} &\sum_{k=1}^N \| \bar{p}^k(\varepsilon_{n+1}^k, \delta_{n+1}) - \bar{p}^k(\varepsilon_n^k, \delta_n) \| \\ &\quad + \| \lambda^k(\varepsilon_{n+1}, \delta_{n+1}) - \lambda^k(\varepsilon_n, \delta_n) \| \\ &\leq C_1 \|\varepsilon_{n+1} - \varepsilon_n\| + C_2 \|\delta_{n+1} - \delta_n\| \\ &\quad + C_3 \|\varepsilon_{n+1} \delta_{n+1}^{-1} - \varepsilon_n \delta_n^{-1}\| := \kappa_n. \end{aligned} \quad (28)$$

*Proof:* The proof is similar to the proof of [34, Th. 2, Sec. 4.3]. ■

## IX. LEARNING ALGORITHM

### A. Regularization, Lagrange Multipliers and Projection Gradient Procedure: The Complete Information Case

In the case when the complete information on the expected payoffs and constraints is available, then the *gradient-like* technique may be applied for the class of diagonal concave games to attain the equilibrium point (see [39])

$$\begin{cases} q_{n+1}^k = p_n^k + \gamma_n^k \nabla_{p^k} L_{\delta_n}^k(p_n^1, \dots, p_n^N; \lambda_n) \\ \mu_{n+1}^k = \lambda_n^k - \gamma_n^k \nabla_{\lambda^k} L_{\delta_n}^k(p_n^1, \dots, p_n^N; \lambda_n) \\ p_{n+1}^k = \pi_{S_{\varepsilon_n}^{N_k}} \{q_{n+1}^k\}, \quad \lambda_{n+1}^{k,l} = [\mu_{n+1}^{k,l}]_+ \\ k = \overline{1, N}, \quad n = 1, 2, \dots \end{cases} \quad (29)$$

where  $\pi_{S_{\varepsilon_n}^{N_k}}\{\cdot\}$  is the projection operator to the simplex set  $S_{\varepsilon_n}^{N_k}$  (9) and  $[\cdot]_+$  is the “positive part” operator. If the parameters of this procedure satisfy  $0 < \gamma_n^k, \delta_n, \varepsilon_n \rightarrow 0$  and  $\sum_{n=1}^\infty \gamma_n^k \delta_n = \infty$ ,  $\sum_{n=1}^\infty \kappa_n < \infty$  [ $\kappa_n$  is defined by (28)], it provides the convergence of the estimates  $p_n^k$  to the corresponding unique equilibrium strategy  $(\bar{p}^1(\vartheta), \dots, \bar{p}^N(\vartheta))$  (see the previous lemma).

### B. Stochastic Approximation Approach and the “Measure Keeping Problem”: Incomplete Information Case

When we deal with incomplete information case, that is, the only current realizations of payoff  $\xi_n^k$  and constraints  $\eta_n^{k,l}$  are available, the “stochastic approximation” version of the procedure (29) can be applied (see, for example, [34]) where instead of  $\nabla_{p^k} L_{\delta_n}^k(p_n^1, \dots, p_n^N; \lambda_n)$  and  $\nabla_{\lambda^k} L_{\delta_n}^k(p_n^1, \dots, p_n^N; \lambda_n)$  their estimates using the current realizations are implemented. Such a procedure is known as the *reinforcement learning algorithm* and can be realized in different ways (with different estimates being used). Next, we present the concrete learning procedure which is based on the LA concept [25], [34].

### C. Bush–Mosteller Reinforcement With Lagrange Multipliers Adjustment

The Bush–Mosteller reinforcement scheme and the normalization procedure presented in [34] will be applied hereafter to design a new learning algorithm for  $N$ -person constrained repeated games with unknown expected payoff and constraints. In fact, we assume that after each stage, the payoff to each player as well as the constraints are random variables. No information concerning the distribution of the payoff and constraints is a

*priori* available. The necessary information is obtained during the course of the game.

*Learning control* is an iterative process involving an adaptation at each stage (time step). We now present the “four-step” recursive algorithm.

*Step 1:* Based on the available data  $u_n^k = u^k(i_k)$ ,  $\xi_n^k$ ,  $\eta_n^{k,l}$ ,  $p_n^k$  ( $p_n^k(i_k) \geq \varepsilon_n^k > 0$ ),  $\lambda_n^k$  build the following functions:

$$\tilde{\xi}_n^k := \xi_n^k - \delta_n p_n^k(i_k) - \sum_{l=1}^{M_k} \lambda_n^{k,l} \eta_n^{k,l} \quad (30)$$

and normalize (scale) them according to the following procedure:

$$\zeta_n^k := 1 - \left( a_n^k \tilde{\xi}_n^k + b_n^k \right) / p_n^k(i_k) \quad (31)$$

where the sequences  $\{a_n^k\}$  and  $\{b_n^k\}$  are defined as follows:

$$\begin{aligned} a_n^k &:= \tilde{a}_n^k \left( \sigma_{\xi,k}^+ + \lambda_n^+ \sum_{l=1}^{M_k} \sigma_{\eta,k,l}^+ + \delta_n \right)^{-1} \\ b_n^k &:= \tilde{a}_n^k + (N_k - 1) (\tau_n^k)^2 [1 + (N_k - 2) \tau_n^k]^{-1} \\ \tilde{a}_n^k &:= \tau_n^k (1 - \tau_n^k) [2(1 + (N_k - 2) \tau_n^k)]^{-1} \\ \tau_{n-1}^k &= \varepsilon_n^k. \end{aligned} \quad (32)$$

*Step 2:* Update the probability distributions  $p_{n+1}^k$  and the Lagrange multipliers  $\lambda_{n+1}$  using the following iterative schemes:  $p_1^k(i) = N_k^{-1}$  ( $i = 1, N_k$ )

$$p_{n+1}^k = p_n^k + \gamma_n^k \cdot \left[ e_{N_k}(u_n^k) - p_n^k + \frac{\zeta_n^k (e_{N_k} - N_k e_{N_k}(u_n^k))}{(N_k - 1)} \right] \quad (33)$$

where

$$e_{N_k}(u_n^k) = \left( \underbrace{0, \dots, 0}_{i_k}, 1, 0, \dots, 0 \right)^T \in R^{N_k}$$

if  $u_n^k = u^k(i_k)$  and  $e^{N_k} = (1, \dots, 1)^T \in R^{N_k}$ . Here, the time-varying correction (adaptation) factors  $\gamma_n^k$  belong to the unit segment. Notice that the conditional mathematical expectation of the normalized environment responses  $\tilde{\xi}_n^k e_{N_k}(u_n^k)$  is equal to the gradient of the augmented Lagrange function with respect to the probability distributions, i.e.

$$E \left\{ \tilde{\xi}_n^k e_{N_k}(u_n^k) \middle| F_{n-1} \right\} \stackrel{a.s.}{=} \frac{\partial}{\partial p_n^k} L_\delta^k(p_n^1, \dots, p_n^N; \lambda_n^k). \quad (34)$$

The Lagrange multipliers are adjusted according to the following recursion:  $\lambda_1^{k,l} > 0$ ,  $\gamma_n^\lambda \geq 0$

$$\lambda_{n+1}^{k,l} = [\lambda_n^{k,l} - \gamma_n^\lambda \psi_n^{k,l}]_0^{\lambda_{n+1}^+} \quad (35)$$

$$\psi_n^{k,l} = \delta_n \lambda_n^{k,l} - \eta_n^{k,l} + b_n^{k,l}. \quad (36)$$

Also note that the conditional mathematical expectation of  $\psi_n^{k,l}$  is equal to the gradient of the augmented (regularized) Lagrange function with respect to the Lagrange multipliers, that is

$$E \left\{ \psi_n^l \middle| F_{n-1} \right\} \stackrel{a.s.}{=} \frac{\partial}{\partial \lambda_n^l} L_\delta^k(p_n^1, \dots, p_n^N; \lambda_n). \quad (37)$$

The operator  $[y]_0^{\lambda_{n+1}^+}$  is defined as follows:

$$[y]_0^{\lambda_{n+1}^+} = \begin{cases} y, & \text{if } y \in [0, \lambda_{n+1}^+] \\ \lambda_{n+1}^+, & \text{if } y > \lambda_{n+1}^+ \\ 0, & \text{if } y < 0. \end{cases} \quad (38)$$

*Step 3:* According to

$$\Pr \{ u_{n+1}^k = u^k(i) \mid F_n \} = p_{n+1}^k(i)$$

generate randomly (for each player) new discrete random variables  $u_{n+1}^k$  as in learning stochastic automata implementation, and get a new observations (realizations)  $\xi_{n+1}^k$  and  $\eta_{n+1}^{k,l}$  that corresponds to the environment vector-reactions.

*Step 4:* Return to Step 1.

The positive sequences  $\{\varepsilon_n^k\}$ ,  $\{\delta_n\}$ ,  $\{\lambda_n^+\}$ ,  $\{\gamma_n^k\}$ , and  $\{\gamma_n^\lambda\}$  will be defined next.

This adaptive algorithm is constructed using the Bush–Mosteller reinforcement scheme (33) with the time-varying correction factors  $\gamma_n^k$ , continuous input  $\tilde{\xi}_n^k$ , and a normalization procedure which is used to ensure the probability measure. It is easy to verify (see [34, Lemma 2, Sec. 4.4]) that  $\zeta_n^k \in (0, 1)$  for any time  $n$ , and  $p_n^k(i) \geq \tau_{n-1} := \varepsilon_n$ ,  $\sum_{i=1}^{N_k} p_n^k(i) = 1$ , that is,  $p_n^k \in S_{\varepsilon_n}^{N_k}$ .

## X. CONVERGENCE ANALYSIS

### A. Main Theorem on the Convergence With Probability One

*Theorem 17 (On Convergence With Probability One):* Suppose that Assumptions A1–A3 hold for the learning reinforcement procedure (33)–(35) and the given game is diagonal convex. In addition, assume that:

- 1) there exist four nonnegative sequences  $\{\gamma_n^k\}$ ,  $\{\gamma_n^\lambda\}$ ,  $\{\delta_n\}$ , and  $\{\varepsilon_n^k\}$  such that  $\gamma_n^k \downarrow 0$ ,  $\delta_n \in (0, \delta^+)$ ,  $\delta_n \downarrow 0$  and  $\varepsilon_n^k \in (0, (1/N_k))$ ,  $\varepsilon_n^k \downarrow 0$ ,  $\limsup_n (\varepsilon_n^k / \delta_n) < \infty$ ;
- 2) the updating factors  $\gamma_n^k$  and  $\gamma_n^\lambda$  are selected as

$$\gamma_n^{\lambda,k} = \frac{\gamma_n^k a_n^k N_k}{(N_k - 1)}, \quad k = \overline{1, N}$$

where  $a_n^k$  is defined by (32) and  $\gamma_n^\lambda$  satisfies  $\sum_{n=1}^\infty \delta_n \sum_{k=1}^N \gamma_n^{\lambda,k} = \infty$ ;

- 3) the following series converges:

$$\sum_{n=1}^\infty \left[ \phi_n + \kappa_n^2 \left( \delta_n \sum_{k=1}^N \gamma_n^{\lambda,k} \right)^{-1} \right] < \infty$$

where

$$\begin{aligned} \kappa_n &:= C_1 \|\varepsilon_{n+1} - \varepsilon_n\| + C_2 |\delta_{n+1} - \delta_n| \\ &\quad + C_3 \|\varepsilon_{n+1} \delta_{n+1}^{-1} - \varepsilon_n \delta_n\| \end{aligned}$$

$$\begin{aligned} \phi_n &:= \kappa_n^2 + C_A^2 \sum_{k=1}^N (\gamma_n^{\lambda,k})^2 + 2C_A \kappa_n \sum_{k=1}^N \gamma_n^{\lambda,k} \\ &\quad + 2M \sum_{k=1}^N (\gamma_n^{\lambda,k})^2 [(\delta_n \lambda_n^+)^2 + (\sigma_\eta^+)^2] \\ &\quad + 2\sqrt{M} \sum_{k=1}^N \gamma_n^{\lambda,k} \kappa_n (\delta_n \lambda_n^+ + \sigma_\eta^+). \end{aligned}$$



Then, the mixed strategies of the players ensure the convergence of the game to the equilibrium point with probability one, that is

$$\sum_{k=1}^N r_k \left( \|p_{n+1}^k - \bar{p}^k(\varepsilon_{n+1}^k, \delta_{n+1})\|^2 + \|\lambda_{n+1} - \lambda(\varepsilon_{n+1}, \delta_{n+1})\|^2 \right) \xrightarrow{a.s.} 0.$$

*Proof:* The techniques used in the proof are based on the Lyapunov approach and Martingale's theory. The fulfilling of **A3** implies that the operating domain of the game is not empty, that is

$$\bigcap_{k=1, N} \bigcap_{l=1, M_k} \{p_n^1, \dots, p_n^N: Q^{k,l}(p_n^1, \dots, p_n^N) \geq b^{k,l}\} \cdot \bigcap_{k=1, N} S_{\varepsilon_n}^{N_k} \neq \emptyset.$$

In view of **A1** and **A2** of this theorem, there exists an integer  $n_0$  such that for any  $n \geq n_0$  the considered game has an equilibrium point defined by  $\bar{p}^k(\varepsilon_n^k, \delta_n)$  and  $\lambda(\varepsilon_n, \delta_n)$ . Consider the following Lyapunov function:

$$W_n(p_n^1, \dots, p_n^N; \lambda_n) := \sum_{k=1}^N r_k \left( \|p_n^k - \bar{p}^k(\varepsilon_n^k, \delta_n)\|^2 + \|\lambda_n^k - \lambda^k(\varepsilon_n, \delta_n)\|^2 \right).$$

Taking into account the adaptive algorithm and, adding and subtracting  $\bar{p}^k(\varepsilon_n^k, \delta_n)$  and  $\lambda^k(\varepsilon_n, \delta_n)$ , respectively, to the first and to the second term of the right side of the previous equality, we obtain

$$\begin{aligned} W_{n+1} \leq \sum_{k=1}^N r_k & \left( \|p_n^k - \bar{p}^k(\varepsilon_n^k, \delta_n) + \gamma_n^k A_n^k \right. \\ & - [\bar{p}^k(\varepsilon_{n+1}^k, \delta_{n+1}) - \bar{p}^k(\varepsilon_n^k, \delta_n)] \|^2 \\ & + \|\lambda_n^k - \lambda^k(\varepsilon_n, \delta_n) - \gamma_n^\lambda \psi_n^k \\ & \left. - [\lambda^k(\varepsilon_{n+1}, \delta_{n+1}) - \lambda^k(\varepsilon_n, \delta_n)] \|^2 \right) \end{aligned}$$

where  $\psi_n := (\psi_n^1, \dots, \psi_n^M)^T$  and

$$A_n^k := e_{N_k}(u_n^k) - p_n^k + \frac{\zeta_n^k(e_{N_k} - N_k e_{N_k}(u_n^k))}{(N_k - 1)}. \quad (39)$$

In the previous upper estimation, we have used the following property of the projection operator:  $|\lambda - [y]_0^{\lambda^+}| \leq |\lambda - y|$ , which is valid for any  $y \in R$ , and  $\lambda \in [0, \lambda^+]$ . Observe that the following bounds hold:  $\|A_n^k\| \leq C_A = \text{const} < \infty$  and  $\|\psi_n^k\| \leq \sqrt{M}(\delta_n \lambda_n^+ + \sigma_\eta^+)$ ,  $\sigma_\eta^+ := \max_l \sigma_{\eta_l}^+$ , that leads to

$$\begin{aligned} W_{n+1}(p_{n+1}^1, \dots, p_{n+1}^N; \lambda_{n+1}) & \leq W_n(p_n^1, \dots, p_n^N; \lambda_n) \\ & + s_n + 2\sqrt{2N} \sqrt{W_n(p_n^1, \dots, p_n^N; \lambda_n) \kappa_n} + \phi_n \end{aligned} \quad (40)$$

where

$$\begin{aligned} s_n := 2 \sum_{k=1}^N & \left( \gamma_n^k (p_n^k - \bar{p}^k(\varepsilon_n^k, \delta_n))^T A_n^k \right. \\ & \left. - 2\gamma_n^{\lambda, k} (\lambda_n^k - \lambda^k(\varepsilon_n, \delta_n))^T \psi_n^k \right). \end{aligned} \quad (41)$$

Based on (34) and (37), perform the following calculations:

$$\begin{aligned} & (p_n^k - \bar{p}^k(\varepsilon_n^k, \delta_n))^T E\{A_n^k | \mathcal{F}_{n-1}\} \\ & = - (p_n^k - \bar{p}^k(\varepsilon_n^k, \delta_n))^T E\{\zeta_n^k e_{N_k}(u_n^k) | \mathcal{F}_{n-1}\} \frac{N_k}{(N_k - 1)} \\ & \stackrel{a.s.}{=} \frac{a_n^k N_k}{(N_k - 1)} (p_n^k - \bar{p}^k(\varepsilon_n^k, \delta_n))^T \frac{\partial}{\partial p_n^k} L_{\delta_n}^k(p_n^1, \dots, p_n^N; \lambda_n) \end{aligned} \quad (42)$$

and

$$\begin{aligned} & (\lambda_n^k - \lambda^k(\varepsilon_n, \delta_n))^T E\{\psi_n^k | \mathcal{F}_{n-1}\} \\ & = (\lambda_n^k - \lambda^k(\varepsilon_n, \delta_n))^T \frac{\partial}{\partial \lambda_n^k} L_{\delta_n}^k(p_n^1, \dots, p_n^N; \lambda_n). \end{aligned} \quad (43)$$

In view of (41)–(43) and *Assumption A2* of this theorem, from the diagonal concavity property (27) it follows:

$$\begin{aligned} & E\{s_n | \mathcal{F}_{n-1}\} \\ & \stackrel{a.s.}{=} 2 \sum_{k=1}^N r_k \left( \frac{\gamma_n^k a_n^k N_k}{(N_k - 1)} \cdot (p_n^k - \bar{p}^k(\varepsilon_n^k, \delta_n))^T \frac{\partial}{\partial p_n^k} L_{\delta_n}^k \right. \\ & \quad \cdot (p_n^1, \dots, p_n^N; \lambda_n) - 2\gamma_n^\lambda (\lambda_n^k - \lambda^k(\varepsilon_n, \delta_n))^T \frac{\partial}{\partial \lambda_n^k} L_{\delta_n}^k \\ & \quad \cdot (p_n^1, \dots, p_n^N; \lambda_n) \Big) \\ & \leq -\gamma_n^\lambda \delta_n \sum_{k=1}^N r_k \left( \|p_n^k - \bar{p}^k(\varepsilon_n^k, \delta_n)\|^2 \right. \\ & \quad \left. + \|\lambda_n^k - \lambda^k(\varepsilon_n, \delta_n)\|^2 \right) \\ & = -\gamma_n^\lambda \delta_n W_n(p_n^1, \dots, p_n^N; \lambda_n). \end{aligned}$$

Substituting this expression into (40) we obtain

$$\begin{aligned} W_{n+1} \leq & (1 - \gamma_n^\lambda \delta_n) W_n(p_n^1, \dots, p_n^N; \lambda_n) \\ & + 2\sqrt{2N} \sqrt{W_n(p_n^1, \dots, p_n^N; \lambda_n) \kappa_n} + \phi_n. \end{aligned}$$

The use of the inequality  $x^r \leq (1-r)x_0^r + (rx_0^{r-1})x$  valid for any  $x$  and  $x_0 > 0$  and any  $r \in [0, 1]$ , for  $r = 1/2$  implies

$$\begin{aligned} 2\sqrt{2N} \sqrt{W_n(p_n^1, \dots, p_n^N; \lambda_n) \kappa_n} & \leq 2\sqrt{2N} \kappa_n \sqrt{x_0} \\ & + \sqrt{2N} \kappa_n (x_0)^{-(1/2)} W_n(p_n^1, \dots, p_n^N; \lambda_n). \end{aligned}$$

By selecting  $\sqrt{x_0}$  such that  $\sqrt{2N} \kappa_n (x_0)^{-(1/2)} = \gamma_n^\lambda \delta_n / 2$ , it follows:

$$\begin{aligned} 2\sqrt{2N} \sqrt{W_n(p_n^1, \dots, p_n^N; \lambda_n) \kappa_n} & \leq 4N \kappa_n^2 (\gamma_n^\lambda \delta_n)^{-1} + \gamma_n^\lambda \delta_n W_n(p_n^1, \dots, p_n^N; \lambda_n) / 2. \end{aligned}$$

Substituting this estimation in the inequality given above, we finally obtain the following upper bound:

$$W_{n+1} \leq (1 - \gamma_n^\lambda \delta_n / 2) W_n + \phi_n + 4N \kappa_n^2 (\gamma_n^\lambda \delta_n)^{-1}. \quad (44)$$

Therefore,  $\{W_n, \mathcal{F}_n\}$  is a quasimartingale. By *Assumption A3* and the Robbins–Siegmund theorem (see [38] or [35, App. ]), the statement of this theorem follows directly from (44). ■

## B. Mean Square Convergence

*Theorem 18 (Mean Square Convergence):* Assume that for the learning reinforcement algorithm, the conditions of the previous theorem, except Condition 3, are fulfilled, and suppose

that  $\phi_n(\gamma_n^\lambda \delta_n)^{-1} + \kappa_n^2(\gamma_n^\lambda \delta_n)^{-2} \rightarrow 0$ . Then the learning algorithm leads asymptotically to the equilibrium of the considered game, in the mean square sense, that is

$$E \left\{ \sum_{k=1}^N r_k \left( \|p_n^k - \bar{p}_n^k(\varepsilon_n^k, \delta_n)\|^2 + \|\lambda_n - \lambda(\varepsilon_n, \delta_n)\|^2 \right) \right\} \rightarrow 0.$$

*Corollary 19:* For the class of the algorithm design parameters defined as follows:

$$\begin{aligned} \gamma_n^k &= \gamma_0^k n^{-\gamma}, \quad \varepsilon_n^k = \varepsilon_0^k n^{-\varepsilon}, \quad \gamma_0^k, \varepsilon_0^k, \delta_0, \lambda_0^+ > 0 \\ \delta_n &= \delta_0 n^{-\delta}, \quad \lambda_n^+ = \lambda_0^+ + n^\lambda \ln n. \end{aligned} \quad (45)$$

Conditions 1–2 of Theorem 1 and the additional condition of Theorem 2 will be verified if

$$\gamma + \varepsilon + \delta + \lambda \leq 1 (\gamma > 0, \varepsilon \geq \delta > 0, \lambda \geq 0) \quad (46)$$

and:

- 1) the convergence with probability one is ensured if

$$2\gamma > 1 \quad (47)$$

- 2) the convergence in the mean square sense is guaranteed if

$$\min\{\gamma - \varepsilon - \delta - \lambda; 1 - 2\delta - \lambda\} > 0. \quad (48)$$

*Proof:* The proof of the convergence with probability one follows directly from the conditions of the previous theorem and in view of the fact that

$$\sum_{n=1}^{\infty} n^{-\alpha} \begin{cases} = \infty, & \text{if } \alpha \leq 1 \\ < \infty, & \text{if } \alpha > 1. \end{cases}$$

Based on [25, Lemma A.5], we obtain the desired result concerning the convergence in the mean square sense. ■

### C. Learning Rate

However, not only the convergence is important but the speed is also essential. For a specific class of the design parameters (45), the next theorem states the convergence rate of the learning game algorithm described above.

*Theorem 20:* Under the conditions of the previous theorems and for the class of the design parameters (45), there exists  $\nu$  such that

$$W_n^* := \sum_{k=1}^N r_k \left( \|p_n^k - \bar{p}_n^k(\vartheta)\|^2 + \|\lambda_n - \lambda(\vartheta)\|^2 \right) \stackrel{a.s.}{=} o_\omega(n^{-\nu})$$

and  $E\{W_n^*\} = o(n^{-\nu})$ , where the order  $\nu$  of the adaptation rate satisfies the following constraint:

$$\begin{aligned} \nu &< \nu^*(\gamma, \delta, \varepsilon, \lambda) \leq \nu^{**} = \frac{1}{3} \\ \nu^*(\gamma, \delta, \varepsilon, \lambda) &= \min\{2\gamma - 1; \gamma + \delta; \varepsilon - \delta + \gamma; 2\delta\} \end{aligned}$$

and the maximum adaptation rate  $\nu^{**} = \nu^*(\gamma^*, \delta^*, \varepsilon^*, \lambda^*)$  is reached for

$$\varepsilon = \varepsilon^* = \delta = \delta^* = \frac{1}{6}, \quad \gamma = \gamma^* = \frac{2}{3}, \quad \lambda = \lambda^* = 0.$$

*Proof:* The proof follows from (44) and the following inequality:

$$\begin{aligned} W_n^* &\leq 2W_n(p_n^1, \dots, p_n^N; \lambda_n) + \sum_{k=1}^N \|p_n^k - \bar{p}_n^k(\vartheta)\|^2 \\ &\quad + \|\lambda_n - \lambda(\vartheta)\|^2 \\ &\leq 2W_n(p_n^1, \dots, p_n^N; \lambda_n) + C\delta_n^2. \end{aligned}$$

The expression of  $\nu^*(\gamma, \delta, \varepsilon, \lambda)$  follows from [34, Lemma A.3-2, App. A]. The optimal design parameters are the solution of the following constrained optimization problem  $\nu^*(\gamma, \delta, \varepsilon, \lambda) \rightarrow \max$  over all the parameters  $\varepsilon \geq \delta$  satisfying (46) and (47). The solution of this problem is achieved when all minimal terms within the operator turn out to be equal, that is

$$\begin{aligned} \nu^{**} &= \max_{(\gamma, \delta, \varepsilon, \lambda)} \min\{2\gamma - 1; \gamma + \delta; \varepsilon - \delta + \gamma; 2\delta\} \\ &\leq (\text{by } \gamma \leq 1 - \delta - \varepsilon - \lambda) \end{aligned}$$

$$\begin{aligned} &\max_{1-\delta-\varepsilon-\lambda} \min\{1 - 2\varepsilon - 2\lambda; 1 - \varepsilon - \lambda; 1 - \lambda - 2\delta; 2\delta\} \\ &= \max_{1-\delta-\varepsilon-\lambda} \min\{1 - 2\varepsilon - 2\lambda; 2\delta\} \\ &= (\text{by } \varepsilon = 1/2 - 2\delta - \lambda) \end{aligned}$$

$$\max_{\delta \leq \varepsilon} 2\delta = 1/3 \text{ under } \delta = 1/6 \text{ and } \lambda = 0.$$

■

## XI. SIMULATION RESULTS

The simulations were realized for the game of three players with the following characteristics of the game:  $N = 3$ ,  $N_k = 2$ ,  $M_k = 1$ , and

$$\begin{aligned} V^1(:, :, 1) &= \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad V^1(:, :, 2) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\ V^2(:, :, 1) &= \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad V^2(:, :, 2) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\ V^3(:, :, 1) &= \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad V^3(:, :, 2) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} q_{(:, 1, 1)}^{1,1} &= q_{(:, 2, 1)}^{1,1} = q_{(:, 1, 2)}^{1,1} = q_{(:, 2, 2)}^{1,1} = [0.25 \quad 0.1] \\ q_{(1, \cdot, 1)}^{2,1} &= q_{(2, \cdot, 1)}^{2,1} = q_{(1, \cdot, 2)}^{2,1} = q_{(2, \cdot, 2)}^{2,1} = [0.1 \quad 0.5] \\ q_{(1, 1, \cdot)}^{3,1} &= q_{(2, 1, \cdot)}^{3,1} = q_{(1, 2, \cdot)}^{3,1} = q_{(2, 2, \cdot)}^{3,1} = [0.5 \quad 0.1] \\ b^{1,1} &= 0.15, \quad b^{2,1} = 0.2, \quad b^{3,1} = 0.8. \end{aligned}$$

The parameters of the learning procedure were as follows:

$$\begin{aligned} \varepsilon = \varepsilon^* = \delta = \delta^* &= 1/6, \quad \gamma = \gamma^* = 2/3, \quad \lambda = \lambda^* = 0 \\ \lambda_0^+ = \delta_0 &= 0.1, \quad \gamma_0 = 0.15, \quad \varepsilon_0 = 0.25. \end{aligned}$$

All sequences maintained their initial values up to time  $n_0 = 300$  and after that, decreased according to the suggested rule. The corresponding learning process is given in Fig. 1 (the current actions probabilities) and Fig. 2 (the averaged payoffs  $\Psi_n^k$ ).

The presented simulations results show the convergence of the mixed strategies as well as the payoff functions to the corresponding values obtained by the implementing of the LP-

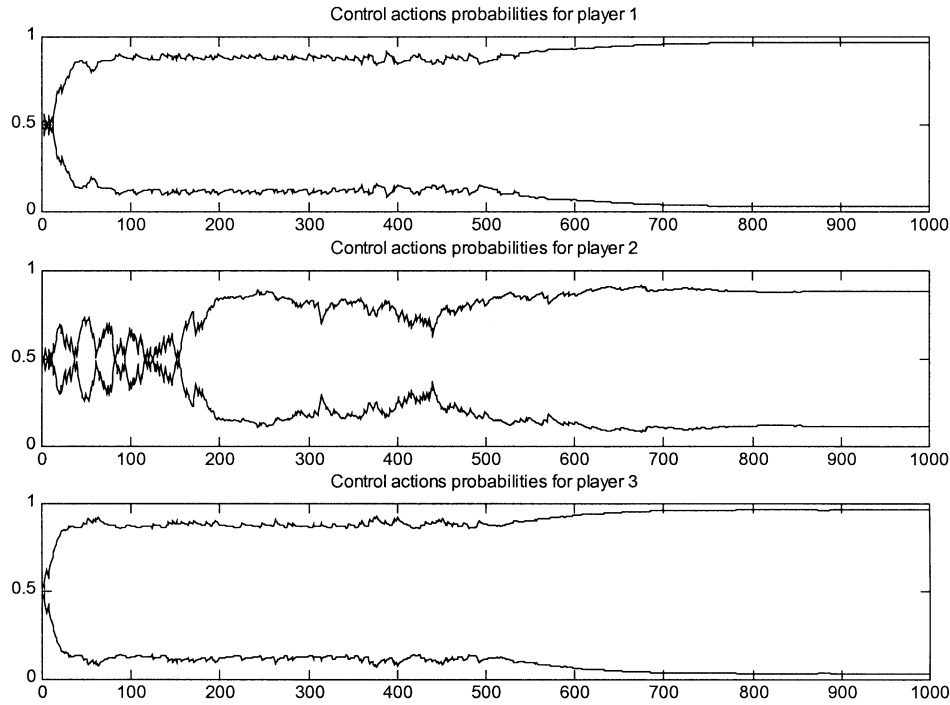


Fig. 1. Control actions probabilities.

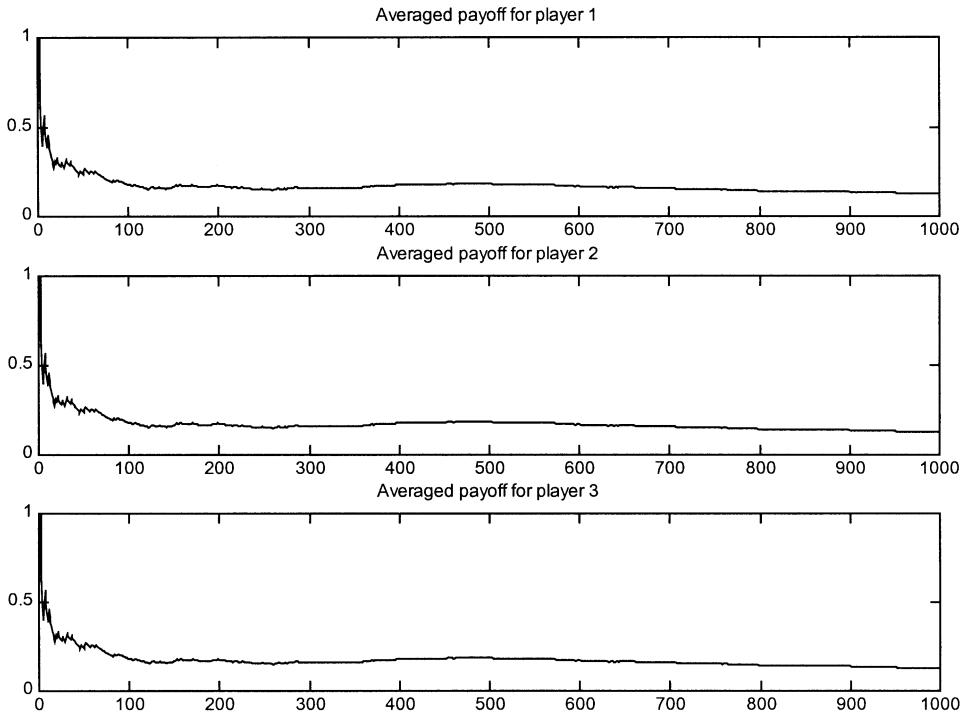


Fig. 2. Averaged payoffs.

approach used for the case of the complete information and described in [9, Ch. 5].

## XII. CONCLUSION

In this paper, a new affine transformation (*normalization procedure*) of current information (payoffs and constraints realizations) is suggested leading to the formation of a scalar

$[0, 1]$ -random variable which is shown to be sufficient for the use in the selected learning reinforcement scheme; this transformation preserves the probability measure for the corresponding mixed strategies over the set of game actions. For each player, the considered reinforcement procedure, based on Lagrange multipliers and an appropriate regularization, is shown to be oriented to obtain the *optimal response* (in average sense) within the given constraints. This learning tactics is proven to *lead to*

the unique Nash equilibrium point if the first moments of the payoffs and constraints realizations form, the so-called, *strictly diagonal concave (SDC) game*. The structure of SDC matrix games is analyzed and it is shown that any diagonal concave matrix game can be presented as the *superposition* of only three of the simplest diagonal concave games: 1) games with nature; 2) two participant zero-sum games; and 3) multiparticipants games where the payoff of each players are independent of his own pure strategies. Based on the stochastic approximation technique, the *convergence* (with probability one and in mean square sense) of the considered learning procedure (using the Bush–Mosteller reinforcement scheme) to the Nash equilibrium is stated and the *rate of learning* is also estimated as  $(n^{-(1/3)})$ .

#### ACKNOWLEDGMENT

The authors are grateful to Dr. E. Gomez Ramirez for the discussion and his help in the simulation process, as well as the anonymous reviewers for their helpful comments and advice.

#### REFERENCES

- [1] E. Altman, *Controlled Markov Decision Processes*. London, U.K.: Chapman & Hall, 1999.
- [2] E. Altman and A. Shwartz, "Adaptive control of constrained Markov chains: Criteria and policies," *Ann. Oper. Res.*, vol. 28, pp. 101–134, 1991.
- [3] J. Arifovic, "Genetic algorithm and the Cobweb model," *J. Econ. Dynam. Contr.*, vol. 18, pp. 2–28, 1994.
- [4] K. J. Arrow, L. Hurwicz, and H. Uzawa, "Constraint qualifications in maximization problems," *Nav. Res. Logist. Q.*, vol. 8, pp. 175–191, 1961.
- [5] N. Baba, *New Topics in Learning Automata: Theory and Applications*. New York: Springer-Verlag, 1984.
- [6] J. F. Mertens and A. Neyman, "Stochastic games," *Int. J. Game Theory*, vol. 10, pp. 53–66, 1981.
- [7] T. E. S. Raghavan and J. A. Filar, "Algorithms for stochastic games—A survey," *ZOR—Methods Models Oper. Res.*, vol. 35, pp. 437–472, 1991.
- [8] T. E. S. Raghavan, T. S. Ferguson, T. Parthasarathy, and O. J. Vrieze, *Stochastic Games and Related Topics*. Norwell, MA: Kluwer, 1991.
- [9] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*. New York: Springer-Verlag, 1979.
- [10] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*. Cambridge, MA: MIT Press, 1998.
- [11] J. P. Aubin, *Mathematical Methods of Game and Economic Theory*. Amsterdam, The Netherlands: North Holland, 1979.
- [12] Y. Balasdo and D. Royer, "Stability of competitive equilibrium with respect to recursive and learning processes," *Dép. d'Econ., Fac. Sci. Econ. Sociales, Univ. Genève, Genève, France*, Rep. 94.03, 1994.
- [13] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, 2nd ed. Philadelphia, PA: SIAM, 1998.
- [14] D. Bertsekas and J. N. Tsiklitis, *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.
- [15] T. Börgers and R. Sarin, "Learning through reinforcement and replicator dynamics," *J. Econ. Theory*, vol. 77, pp. 1–14, 1997.
- [16] X. Chen and H. White, "Nonparametric adaptive learning with feedback," *J. Econ. Theory*, vol. 82, pp. 190–222, 1998.
- [17] M. Duflo, *Random Iterative Models*. New York: Springer-Verlag, 1997.
- [18] J. Dutta and K. Prasad, "Learning by observation within the firm," *J. Econ. Dynam. Contr.*, vol. 20, pp. 1395–1425, 1996.
- [19] Y. M. El-Fattah and C. Foulds, *Learning Systems: Decision, Simulation, and Control*. New York: Springer-Verlag, 1978.
- [20] I. Erev and A. E. Roth, "Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria," *Amer. Econ. Rev.*, vol. 88, pp. 848–881, 1998.
- [21] D. Friedman, "Evolutionary economics goes mainstream: A review of the theory of learning in games," *J. Evol. Econ.*, vol. 8, pp. 423–432, 1998.
- [22] E. Kalai and E. Lehrer, "Rational learning leads to Nash equilibrium," *Econometrica*, vol. 61, pp. 1019–1045, 1993.
- [23] C. M. Kuan and H. White, "Adaptive learning with nonlinear dynamics driven by dependent processes," *Econometrica*, vol. 62, pp. 1087–1114, 1994.
- [24] S. Lakshmivarahan, *Learning Algorithms Theory and Applications*. New York: Springer-Verlag, 1981.
- [25] K. Najim and A. S. Poznyak, *Learning Automata: Theory and Applications*. New York: Pergamon, 1994.
- [26] ———, "Multimodal searching technique based on learning automata with continuous input and changing number of actions," *IEEE Trans. Syst., Man, Cybern. B*, vol. 26, pp. 666–673, Aug. 1996.
- [27] K. Najim, A. S. Poznyak, and E. Gomez, "Adaptive policy for two finite Markov chains zero-sum stochastic games with unknown transition matrices and average payoffs," *IFAC Automatica*, vol. 37, pp. 1008–1018, 2001.
- [28] K. S. Narendra and M. A. L. Thathachar, *Learning Automata: An Introduction*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [29] J. Nash, "Equilibrium points in  $n$ -person games," *Proc. Nat. Acad. USA*, vol. 36, pp. 48–49, 1950.
- [30] A. V. Nazin and A. S. Poznyak, "Matrix  $N$ -person game with incomplete information," *Econ. Math. Meth.*, vol. 14, pp. 958–968, 1978.
- [31] ———, *Adaptive Choice of Variants* (in Russian). Moscow: Nauka, 1986.
- [32] M. F. Norman, *Markov Processes and Learning Models*. New York: Academic, 1972.
- [33] S. Özyildirim and N. M. Alenmdar, "Learning the optimum as a Nash equilibrium," *J. Econ. Dynam. Contr.*, vol. 24, pp. 483–499, 2000.
- [34] A. S. Poznyak and K. Najim, *Learning Automata and Stochastic Optimization*. New York: Springer-Verlag, 1997.
- [35] A. S. Poznyak, K. Najim, and E. Gomez, *Self-Learning Control for Finite Markov Chains*. New York: Marcel Dekker, 2000.
- [36] A. Poznyak and K. Najim, "Bush–Mosteller learning for zero-sum repeated game with random pay-offs," *Int. J. Syst. Sci.*, vol. 32, no. 10, pp. 1251–1260, 2001.
- [37] A. E. Roth and I. Erev, "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term," *Games Econ. Beh.*, vol. 8, pp. 164–212, 1995.
- [38] H. Robbins and D. Siegmund, "A convergence theorem for nonnegative almost supermartingales and some applications," in *Optimizing Methods in Statistics*, J. S. Rustagi, Ed. New York: Academic, 1971.
- [39] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave  $N$ -persons games," *Econometrica*, vol. 33, pp. 520–534, 1965.
- [40] P. S. Sastry, V. V. Phansalkar, and M. A. L. Thathachar, "Decentralized learning of Nash equilibria in multiperson stochastic game with incomplete information," *IEEE Trans. Syst., Man, Cybern.*, vol. 24, pp. 769–777, May 1994.
- [41] N. Shimkin, "Stochastic games with average cost constraints," in *Annals of the International Society of Dynamic Games*, T. Basar and H. Haurie, Eds. Cambridge, MA: Birkhäuser, 1994, vol. 1, Advances in Dynamic Games and Applications.



**Alexander S. Poznyak** (M'97) was born in Moscow, Russia, on December 6, 1946. He received the M.S. degree from Moscow Physical Technical Institute (MPhTI), Moscow, in 1970, and the Ph.D. and Dr. degrees from the Institute of Control Sciences of the Russian Academy of Sciences, Moscow, in 1978 and 1989, respectively.

From 1973 to 1993, he was a Researcher and Leading Researcher at the Institute of Control Sciences. Since 1993, he has been a Full Professor at CINVESTAV-IPN, Mexico City, Mexico. He has published more than 100 papers in different international journals and seven books including *Adaptive Choice of Variants* (Moscow, Russia: Nauka, 1986), *Learning Automata: Theory and Applications* (New York: Elsevier-Pergamon, 1994), *Learning Automata and Stochastic Programming* (New York: Springer-Verlag, 1997), *Self-Learning Control of Finite Markov Chains* (New York: Marcel Dekker, 2000), and *Differential Neural Networks: Identification, State Estimation and Trajectory Tracking* (Singapore: World Scientific, 2001). He is also Associate Editor of CDC and ACC.

Dr. Poznyak is a member of the Mexican Academy of Sciences and System of National Investigators (SNI-3). He is also a Member of the editorial board of IEEE Control Systems Society.



**Kaddour Najim** was born in Aïn Sbit, Morocco, on June 22, 1949. He received the Ph. degree in automatic control from the University Paul Sabatier, Toulouse, France, in 1979.

Since 1982, he has been Full Professor at the National Polytechnic Institute of Toulouse (ENSI-ACET-INPT). His main research interests are in the areas of process modeling and control, adaptive control, nonlinear systems modeling and control, neuro-fuzzy systems, learning stochastic automata, stochastic optimization, adaptive control of Markov chains, development of adaptive strategies for stochastic games, and reliability. He is the author or coauthor of more than 14 books and 100 papers. He serves on the editorial board of *Journal of Systems Engineering*.